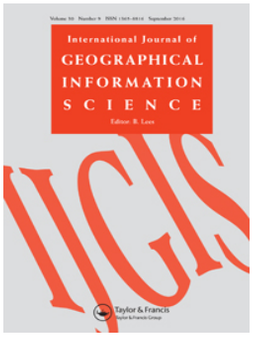

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: http://www.tandfonline.com/loi/tgis20

A framework for identifying activity groups from individual space-time profiles

Jianan Shen & Tao Cheng

To cite this article: Jianan Shen & Tao Cheng (2016) A framework for identifying activity groups from individual space-time profiles, International Journal of Geographical Information Science, 30:9, 1785-1805, DOI: 10.1080/13658816.2016.1139119

To link to this article: http://dx.doi.org/10.1080/13658816.2016.1139119

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE, 2016
VOL. 30, NO. 9, 1785–1805
http://dx.doi.org/10.1080/13658816.2016.1139119

Taylor & Francis
Taylor & Francis Group

A framework for identifying activity groups from individual space-time profiles

Jianan Shen and Tao Cheng

SpaceTimeLab for Big Data Analytics, Department of Civil, Environmental and Geomatic Engineering, University College London, London, UK

ABSTRACT

Datasets collecting the ever-changing position of moving individuals are usually big and possess high spatial and temporal resolution to reveal activity patterns of individuals in greater detail. Information about human mobility, such as 'when, where and why people travel', is contained in these datasets and is necessary for urban planning and public policy making. Nevertheless, how to segregate the users into groups with different movement and behaviours and generalise the patterns of groups are still challenging. To address this, this article develops a theoretical framework for uncovering space-time activity patterns from individual's movement trajectory data and segregating users into subgroups according to these patterns. In this framework, individuals' activities are modelled as their visits to spatio-temporal region of interests (ST-ROIs) by incorporating both the time and places the activities take place. An individual's behaviour is defined as his/her profile of time allocation on the ST-ROIs she/he visited. A hierarchical approach is adopted to segregate individuals into subgroups based upon the similarity of these individuals' profiles. The proposed framework is tested in the analysis of the behaviours of London foot patrol police officers based on their GPS trajectories provided by the Metropolitan Police.

ARTICLE HISTORY
Received 3 July 2015
Accepted 30 December 2015

KEYWORDS
Activity pattern; region of interest; behaviour similarity; clustering; time geography

1. Introduction

Things people do in space and time have long been a research topic in behavioural and socio-economic studies, with particular focus on the highly dynamic urban environment (Cullen 1972, Chapin 1974). The term 'activity pattern' in this research is used to describe patterned ways in which groups of people carry out their daily activities. These activities are naturally linked to the places where they are undertaken and the times (e.g. time of day, day of week or year) at which they take place. By segregating communities into groups of people sharing similar activity patterns, many socio-economic and socio-demographic problems and their ties with individual behaviour preferences can be revealed (Chapin 1974). Research into these patterns attempts to

CONTACT Jianan Shen jianan.shen.13@ucl.ac.uk

This article was originally published with error. This version has been amended. Please see Corrigendum (http://dx.doi.org/10.1080/13658816.2016.1151332).

answer questions about the life styles, behaviours, routines and preferences of different groups of people.

Early studies of human activity patterns were confined to traditional statistical and survey studies because of a lack of large-scale activity data and the tools to deal with them. Nowadays, thanks to the ubiquity of telecommunication and sensor technologies, such data are now available in the form of GPS trajectories and mobile phone user data at decreasing cost. Movements are continually recorded as trajectories, which are sequences of geo-located and time-stamped points, often with associated information (Kuijpers and Vaisman 2007). GPS, mobile phone service and location-based app data are typical examples of these new datasets. They are often big and possess high spatial and temporal resolutions, which enable researchers to explore movement patterns in greater detail than before.

Most current research trying to make use of this kind of data for behavioural analysis focuses on the spatial or temporal dimensions in isolation (Timmermans *et al*. 2002, Kwan 2004, Andrienko *et al*. 2011). Li *et al*. (2008) uses space and place as a depiction of human activity patterns, while Wilson (2001, 2007) analyses human activities in time based on duration, and time sequences. However, these studies neglect the fact that space and time play equally significant roles in the description of people's activities and therefore do not provide a complete indication of people's activity patterns. In reality, people carry out different activities at different places at different times of the day. The activity they are doing is not only indicated by where they are, but also how long they spend in the place and when they do it. This is also because time is a resource, how people allocate the length of their time resource on particular activities also varies (Szalai 1966).

This article aims to build on previous work that views the spatial and temporal domains in isolation and establish a universal framework that enables comprehensive analysis of space and time in order to group people with similar behaviour patterns based upon trajectory data. The framework segregates individuals into subgroups based upon where (place), when (time) and how long (duration) the activities are conducted for each individual. The next section will present the related work and methodological framework. The methods in each step of the framework are explained in Section 3. To demonstrate its capability, this framework is tested using officer foot patrol dataset provided by the London Metropolitan Police presented in Section 4, and Section 5 summarises the major findings and directions for further research.

2. Methodological framework

2.1. *Related works*

The work proposed in this article is essentially rooted in the pioneering work of Chapin (1974), where the author introduced the measure of activities by allocation of 'time budget' and surveys how people from different socio-economic background spend their time in different places carrying out activities. Following Chapin's ideas, Cao *et al*. (2010) and Li *et al*. (2008) introduced models for semantic movement trajectories, by using new movement datasets made available by the technological advances of GPS. In their research, the trajectories are defined as sequences of stops and movements from

place to place with time tags and semantic meanings in the geographic background. Similar to Palma's assumption (Palma *et al*. 2009), the logic behind this is that the place the user stays and the time when the user stays indicates the interest to her/him. Therefore, they all define the region where multiple users stop as their common region of interests (ROIs).

Based on the ROI generated along the movement trajectories, semantic analysis of movements has been introduced in Alvares *et al*. (2007) and Cao *et al*. (2010). Subsequent research enriched the semantic meaning of visited places along the raw trajectories to interpret high-level behaviours and activities (Baglioni *et al*. 2009). Based on the visiting pattern of different places, Zheng *et al*. (2009) defined a similarity metric to detect and group similar users. Temporal patterns have also been used to distinguish users and places. Andrienko *et al*. (2015) used the temporal patterns in which groups of users visit different places to discover the semantics without the invasion of privacy. More commonly, sequence alignment methods have been widely used in the temporal analysis of movements (Shoval and Isaacson 2007, Delafontaine *et al*. 2012, Kwan *et al*. 2014). An *et al*. (2015), as well as Long and Nelson (2013), made comprehensive reviews on the paradigm shift from place-based analysis of movement data to time geography and further into new space-time methods and models.

From a spatial point of view, the concept of 'where you stop is who you are', proposed by Spinsanti *et al*. (2010), posits that individuals' activities are associated with places. Progress has been made in defining movement patterns with series of semantic locations according to users' travel sequences, which can be used to group users based on location-based similarity metrics (Li *et al*. 2008, Xiao *et al*. 2010, Mckenzie 2014).

From a temporal point of view, the concept of 'what you are is when you are', proposed by Ye *et al*. (2011) applies temporal activeness profiles to define the similarity between check-in activities in location-based social networks. Such temporal profiles have also been applied to quantify the description of human mobility and for behaviour similarity analysis (Jankowski *et al*. 2010, Vazquez-Prokopec *et al*. 2013, Andrienko *et al*. 2015). In these approaches, hierarchical clustering (HC) (Ward *et al*. 1963) methods are mainly used to segregate different users because of their common adoption in taxonomy. Since HC uses similarity matrices as inputs, researchers can define their own distance or similarity metrics to generate a similarity matrix for HC according to the research purpose.

2.2. *Methodological framework*

Extending from the ideas of indicating activity patterns from spatial locations ('where you stop is who you are') and temporal activeness ('what you are is when you are'), as well as taking Chapin's idea (Chapin 1974) of using time as an accounting device to study social behaviours, we propose the concept that 'where, when and how long you stay is who you are'. This means that if we can generate profiles describing 'where', 'when' and 'how long' individuals undertake certain activities, we can group people with similar profiles and segregate them into subgroups with different behaviours and look at their relationships with socio-economic factors. For such a purpose, the proposed framework should be capable of:

Figure 1. Flow chart of the framework.

- Discovering interesting places in space and time, telling where the places are and when they are 'interesting';
- Describing individual profiles with 'where', 'when' and 'how long';
- Measuring behavioural differences based on the profiles;
- Grouping people with similar behavioural patterns;
- Explaining how these subgroups are formed.

These five points are realised by five steps based upon movement trajectory data as follows:

(1) **ST-ROI detection**: Extracting ST-ROI, i.e. the ROIs with not only spatial location information but also time spans when they are 'interesting';
(2) **Individual space-time profiling**: Simplifying users' movement patterns as one-dimensional ST-ROI visit sequential patterns, i.e. the patterns of people visiting and leaving different places at different times. Defining individual's time allocation on the ST-ROIs as their 'space-time profiles' to depict his/her activity routine;
(3) **Pair-wise profile comparison**: Defining a new similarity metric of activities with the similarity of time allocation profiles;
(4) **Hierarchical clustering**: Using the proposed similarity metrics in the clustering analysis to segregate people into groups of different activity patterns;
(5) **Semantic validation**: Validating the performance of the HC method and explaining the generated user subgroups by taking geographic background and event records into account.

These five steps are demonstrated by the flow chart in Figure 1.

The next section is organised according to the procedure illustrated in the flow chart, explaining the method developed in each step of the framework.

3. Steps from trajectory data to segregated groups

3.1. *ST-ROI detection*

In the study of the movement of individual users, the basic assumption in many existing works is that people stop at a certain place to undertake various activities and leave for the next place. Therefore, the stopping behaviour is of greater interest than the moving for the detection of interesting places (Palma *et al*. 2009). Identifying stops in the trajectories is the first step in these researches (Alvares *et al*. 2007, Zheng *et al*. 2009). For movement datasets that possess relatively high and regular sampling rates, such as GPS tracking data, the point where an individual stops moving or moving slower than

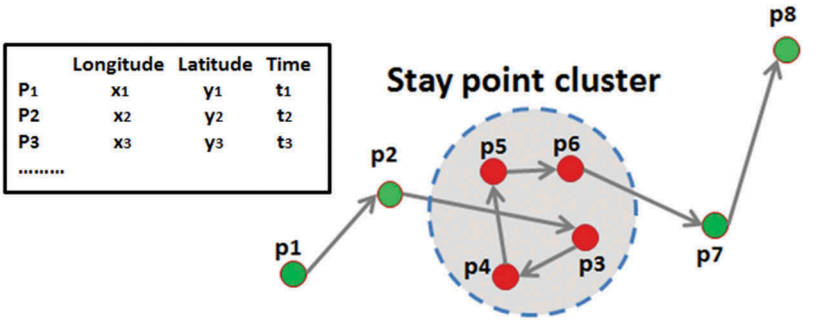

Figure 2. The GPS trajectory of a user (Zheng *et al*. 2009). The green nodes represent non-stay points where the user directly passes by, and the red ones represent stay points.

the pre-defined speed threshold (e.g. lower than 5 km per hour for police foot patrol officers) can be considered to be a stay point. Figure 2 is a typical example proposed by Zheng *et al*. (2009) of how activities are represented by the stay points and the movement parts that connect stay points.

Because not all stops are considered interesting in pedestrian behaviours in urban environment, the spatio-temporal region of interest (ST-ROI) is defined as the place with frequent visit of multiple users in a limited time span. In other words, an ST-ROI is a region of high density clustering of stay points with spatial boundaries, as well as start and end times.

Several density-based methods have been used for discovering ROIs, but most of them are used to aggregate spatial point objects (Cao *et al*. 2010, Lee *et al*. 2013). Li introduced OPTICS to take the advantages of both HC and density-based clustering to detect ROIs in multiple scales (Li *et al*. 2008). The widespread use of density-based clustering methods in ROI detection is because the working mechanism of density-based clustering (DBSCAN) enables it to detect clusters of arbitrary shapes without specifying the number of clusters in the data a priori. It also has a notion of noise and is tolerant to outliers. Moreover, because the algorithm can work directly with a database, the clustering process can be sped up by optimising query strategy in the database (Patwary *et al*. 2012).

Among the many variations of density-based approach to cater to different research purposes, ST-DBSCAN (Birant and Kut 2007) is an extension particularly developed to deal with space and time intervals comprehensively. Besides the advantages inherited from DBSCAN, ST-DBSCAN has features of its own to make it even more effective for detecting ST-ROIs. By extending the idea of traditional DBSCAN, the ST-DBSCAN not only sets up maximum reachable distance (MRD) in space but also in time. Any stay point must satisfy the criteria of spatial MRD and temporal MRD simultaneously to be included in the spatio-temporal cluster. While other density-based methods can only use one MRD parameter for all types of variables no matter whether they share same measurement units or not, ST-DBSCAN enables us to set spatial and non-spatial (temporal) MRD separately according to the nature of the moving data we are working on.

ST-DBSCAN is capable of clustering objects with a combination of both spatial and temporal measurements and detecting noise when different densities exist. These

characteristics make ST-DBSCAN the best option to detect the location as well as the life span of ST-ROIs, revealing where ST-ROIs are, when they emerge and when they disappear. As far as the semantic meaning of the place is concerned, the life span of ST-ROIs are of equal importance as their locations because interesting places are not always busy all day long and can become interesting for different reasons in different time periods. Therefore, it is possible for ST-DBSCAN to find places that are interesting for different groups at different times of the day. A short street segment with bars and an underground station nearby, for instance, can be busy in morning peak because of the commuters' intensive visit to the underground station and then become lively again at midnight when London Undergrounds stop service and bars reach their business climaxes for the relaxing people. The ST-ROIs can be visualised in a space-time cube (Andrienko *et al*. 2010) as shown in Figure 3. It can be seen that although the town centre is an interesting place, it does not always draw the officers' attention throughout the entire day.

As to determining the parameters in density-based clustering, many researchers have optimised the parameters by tuning according to the domain knowledge and aided by visual representation (Cortez *et al*. 2014). ST-DBSCAN has three parameters. They are SMRD, TMRD and MinPts (the minimum number of reachable points needed to form a new cluster). In a similar way, we first determine SMRD and TMRD according to the estimated GPS spatial error and time resolution in the automatic personnel location systems (APLS) dataset. Then the MinPts is defined in Equation (1), determined by calculating the neighbourhood of every point in dataset as proposed by Zhou *et al*. (2012).

$$\text{MinPts} = \frac{1}{n}\sum_{i=1}^{n} p_i \tag{1}$$

where p_i is the number of points in SMRD and TMRD neighbourhood of point i, and n is the total number of all the points.

To ensure the speed of processing large point datasets, a kd-tree (Wald and Havran 2006) for space index acceleration is used in the ST-DBSCAN algorithm to optimise the neighbour searching strategy.

3.2. *Individual space-time profiling*

Knowing the ST-ROIs and the stops and moves of every individual user, we can establish a model in which the time and spatial aspects are considered in a joint effort. We describe an individual's moving process by noting when the user visits a particular ST-ROI and how long she/he stays before leaving for another ST-ROI. In this way, the movement description of a user can be represented by the time she/he arrives at an ST-ROI and then leaves for another; thus the whole movement of a user in a day is simplified as a series of ST-ROIs she/he visits. Figure 3 shows how two persons' trajectories (Figure 3a) are simplified as two sequences of their respectively visited ST-ROIs (Figure 3b) with the time information recording when they arrive at and leave each ST-ROI. This simplified representation is also widely used in movement pattern studies (Zheng 2011) but without taking the time span of interesting places into account.

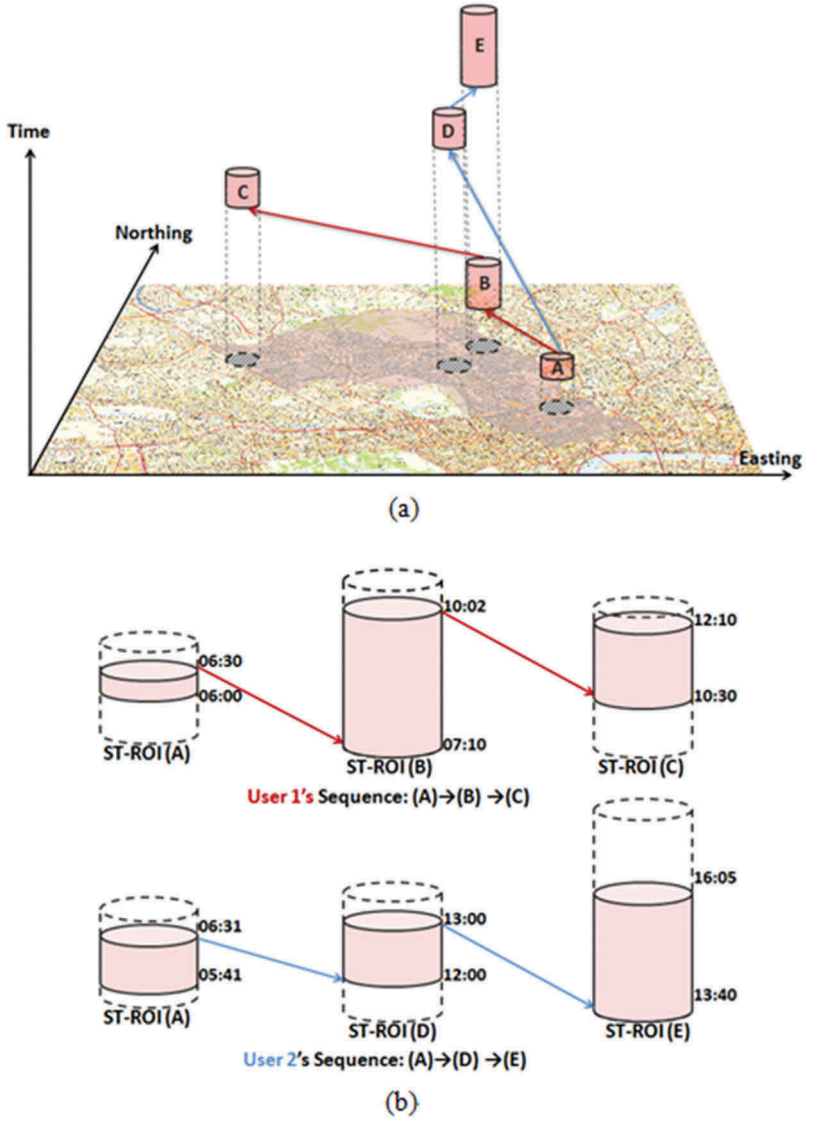

Figure 3. The simplified representation of two example users' movements (a) with the trajectory of two users in space-time; (b) simplified movements with sequence of time-stamped ST-ROIs.

Based on this simplified representation of individual movements, the daily behaviour routine of individuals in the study period can be expressed by how much time each user spends in different ST-ROIs. As the example shown in Figure 3, A, B, C, D and E are the major ST-ROIs frequently visited by two users. The circular shadow of the ST-ROIs projected on the base map indicates their spatial locations. It can be noticed that B and E are spatially located at the same place but not at the same time. In this example, user 1 keeps active from 06:00 to 12:10 in the day. She/he spends approximately 0.5 hour, 3 hours and 2 hours in ST-ROI (A), ST-ROI (B) and ST-ROI (C), respectively, while

user 2 spends about 1 hour, 1 hour and 2.5 hours of the stopping time in ST-ROI (A), ST-ROI (D) and ST-ROI (E), respectively, from 05:41 to 16:05.

One of the advantages of this model over the purely spatial models is that it considers not only spatial information but also temporal factors so that more information can be discovered. In our model, although two ST-ROIs may be in the same spatial location, they can exist in different time periods with different time spans and have a clear temporal gap in between with no activity linking these two spatio-temporal entities as one. Therefore, the semantic meaning of these two ST-ROIs may be different. Taking Figure 3(a) as an example, ST-ROI (B) and (E) are in the same location, but user 1 visits ST-ROI (B) in the morning and user 2 visits ST-ROI (E) at night. ST-ROI (B) and (E) locate at the same place but the purposes of visits can still be very different because of the differences in time.

3.3. *Profile comparison – dissimilarity measure*

In terms of similarity of activity patterns, it is assumed that individuals usually stop at places for certain objectives. Different social groups may have different preferences and habits that may lead to dissimilarities in their movement patterns and reactions to certain events (Chapin 1974). Based on the pattern in which individuals stop at a series of places, various similarity metrics are proposed with emphases on different features of movements.

The most commonly used similarity metrics are the similarity considering the sequences of ROI visited. Different methods such as the longest common subsequence (LCS) (Dodge *et al*. 2009), multiple sequence alignment (Kwan *et al*. 2014) and edit distance (Chen *et al*. 2005), and trajectory clustering (Nara *et al*. 2011) have been used for measuring similarity in terms of sequence relationships.

In our study, the stay durations and time distributions are the major concern. However, the information of sequences in which an individual visits different places can still be preserved. This is because the generated ST-ROIs include start and end times and people can only visit an ST-ROI that exists early in the day first before visiting an ST-ROI that comes later. For example, a user can start his day from a coffee shop that is of great interest to lots of people in the early morning and then go to work in a business centre that become 'interesting' afterwards.

The similarity of movement patterns was defined in other studies by the common visited places. A typical expression of the place-based similarity between user 1 and user 2, as proposed by Zheng *et al*. (2009) in GeoLife2.0, is calculated as follows:

$$\mathrm{SIM}_{\mathrm{user}}(1,2) = \frac{\sum_{p \in \mathrm{ROIS}_{1,2}} \frac{1}{F_p}}{\sqrt{\left(\sum_{p \in \mathrm{ROIS}_1} \frac{1}{F_p}\right) * \left(\sum_{p \in \mathrm{ROIS}_2} \frac{1}{F_p}\right)}} \tag{2}$$

Here, $\mathrm{ROIS}_{1,2}$ is the set of places visited by both user 1 and 2, while ROIS_1 and ROIS_2 represent the sets of places visited by user 1 and 2, respectively. F_p is the popularity index of these places, and it is calculated according to the number of individuals that have visited there. The popularity indices are used as the denominator in the weight attached to different places. Weighting different places allows for a case in which two

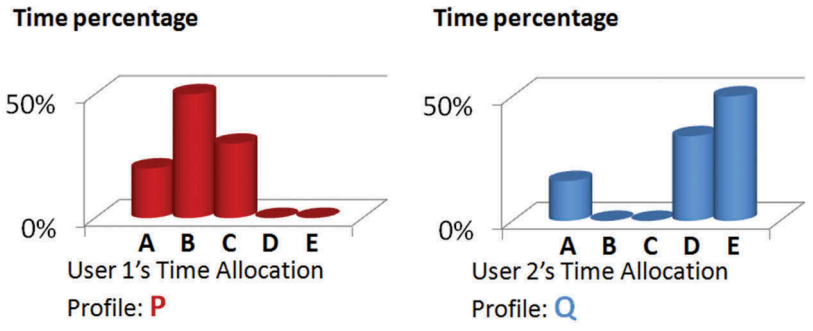

Figure 4. Histogram showing the percentage of the time two users allocate to ST-ROIs.

users go to a common place that is visited by many other users. In such a case, the impact on the behavioural similarity of this place should be smaller than those of places that have been visited by users 1 and 2, but which are not usually visited by other individuals.

In our proposed model, the basic assumption is that people of different socio-economic compositions allocate time to different places and phases of the day for different pursuits in their everyday affairs. Not only the places, but also the time of the activity indicates the behavioural preference of the individual. For example again, in Figure 3(a), B and E are ST-ROIs that are geographically located in the same place, but the reason why people visit them can differ at different times of the day.

With the model describing the behaviours of individuals as movements from one ST-ROI to another, the patterns of how users spend different percentages of their time in each of the ST-ROIs are acquired. Just like research that uses time allocation to indicate personal characteristics in behavioural studies (Kölbl and Helbing 2003), we use the profiles of time allocation in ST-ROIs (Figure 4), called 'space-time profiles', as a measure of activity features. The question now is how to quantify the pair-wise similarity of the movement patterns based on these space-time profiles so that they can be used as a defined distance metric in the following clustering analysis. To satisfy the requirements of clustering analysis as well as the purpose of the behaviour comparison, discrete Jensen–Shannon divergence (JSD) (Lin 1991) is used to measure the dissimilarity of the time distribution profiles of two users.

Classic information theory concepts have the potential to be applied to new space-time data (Tsou 2015). JSD, as demonstrated in Equation (4), is also known as the information radius. It is a popular method used in information theory and taxonomy in bioinformatics, measuring the dissimilarity of multiple probability distributions. JSD is an extension of the Kullback–Leibler divergence (KLD) (Kullback and Leibler 1951), which is based on Jensen's inequality and the Shannon entropy. Some remarkable ameliorated properties of JSD make it especially suitable for our research:

(1) Unlike the well-known Kullback divergences, JSD does not require the condition of absolute continuity of the distributions. It can be applied to discrete distributions just like the space-time profile shown as the percentage histogram in Figure 5.
(2) Unlike many other similarity metrics used in information theory, the JSD between two distributions P and Q is symmetric, which means that JSD(P, Q) is equal to JSD

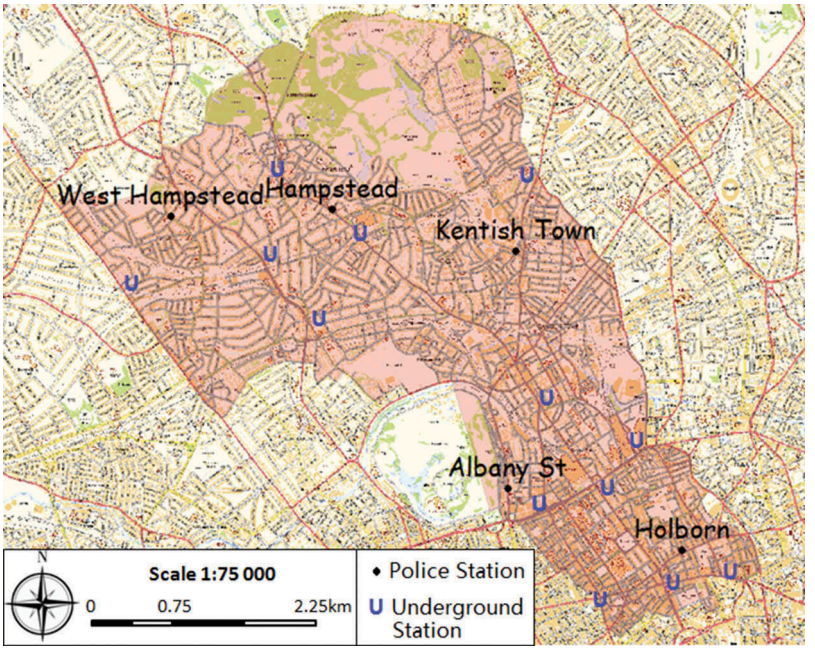

Figure 5. Police stations and Underground stations in the Borough of Camden.

(*Q*, *P*). This symmetric characteristic is similar to the distances between objects, which enables it to act as a distance metric in clustering analysis.

(3) The upper bound of the JSD has been proven to be no greater than 1 (Lin 1991). These bounds are crucial for the definition of similarity.

The following equations show how the discrete version of the JSD is derived from the KLD. According to this equation, the JSD ranges between 0 and 1. The closer the JSD is to 1, the larger the difference between the two space-time profiles.

$$\mathrm{KLD}(X||Y) = \sum_i X(i) \ln \frac{X(i)}{Y(i)} \tag{3}$$

where X and Y are two distributions to be compared by the KLD and $X(i)$ is the i-th term in the distribution X.

$$\mathrm{JSD}(P||Q) = \frac{1}{2}\mathrm{KLD}(P||M) + \frac{1}{2}\mathrm{KLD}(Q||M)$$

where P and Q are the two users' space-time profiles, $M = \frac{1}{2}(P + Q)$.

$$\mathrm{JSD}(P||Q) = \frac{1}{2}\sum_i P(i) \ln \frac{2 \cdot P(i)}{P(i) + Q(i)} + \frac{1}{2}\sum_i Q(i) \ln \frac{2 \cdot Q(i)}{P(i) + Q(i)} \tag{4}$$

Whenever $P(i) = 0$, the contribution of i-th term to JSD is interpreted as 0 because $\lim_{x \to 0} x \ln(x) = 0$

3.4. *Hierarchical clustering*

With the JSD-based similarity metric (Equation 4), a dissimilarity matrix can be calculated. Each element in the matrix represents the pair-wise dissimilarity of two users' profiles. This pair-wise dissimilarity matrix can be processed by a HC algorithm for user segregation. The strength of HC is that any valid measure of distance can be used, including self-defined distance metrics. Furthermore, the observations themselves are not required: all that is used is a matrix of pair-wise distances.

The number of clusters to be generated can be determined by the Dunn index (DI_m) (Dunn 1973) that quantifies how well the dataset is separated. As defined in Equation (5), the Dunn index is the ratio between the minimal inter-cluster distance of m clusters to the maximal intra-cluster distance in each cluster:

$$DI_m = \frac{\min\limits_{1 \le i < j \le m} \delta(C_i, C_j)}{\max\limits_{1 \le k \le m} \Delta k} \tag{5}$$

The Dunn index is also chosen as an evaluation metric to compare the HC results with other clustering methods in grouping users as shown in the case study (Section 4.3). Grouping results can be demonstrated as a taxonomy tree (as shown in Figure 7) in a hierarchical structure. The results can also be visualised in space-time cubes to provide a more intuitive sense of difference (as shown in Figure 10).

However, the HC is time-consuming. When the number of users increases, the computation intensity of HC increases at an exponential growth rate. Therefore, a parallel implementation of HC can be used to significantly speed up the grouping process (Rajasekaran 2005).

3.5. *Semantic validation*

The Dunn index compares the differences within and across clusters, which can provide some insights in determining the convergence and appropriate number of clusters in a relative sense. However, it does not actually validate the clustering results. Therefore, we also use independent and ground truth data source, including building type and officer type information, to support the semantic validation of grouping results as shown in Section 4.4.

4. Case study

4.1. *Data introduction*

The study took place in the Camden Borough (Figure 5), which lies to the north of central London, United Kingdom. Five major police stations are located in this region, namely, West Hampstead, Hampstead, Kentish Town, Albany Street and Holborn. The research was centred on the police foot patrol activities within Camden.

The major dataset captured officers' location stamps recorded by GPS-integrated portable radio sets that were carried by every officer in the field and are uploaded to the APLS of the Metropolitan Police for operational use. The chosen study period

covered February 2012. All 84,027 records generated by 355 officers provided call sign information, device IDs, as well as the time sequence and locations. Usually, the data were logged every 10 minutes, which is a much lower temporal resolution than one would normally expect from GPS data. A record was also inserted when an officer called with his/her radio. One call sign can only be used by one officer and it cannot be changed until she/he leaves his/her present unit. Thus, it can be assumed that one call sign uniquely represents one police officer. In the practical application, the APLS data are not perfect. Many records have been lost and not every move of every officer was logged into the system for various reasons such as device power-off and signal failure. Therefore, to ensure that the GPS log reliably reflected where officers truly were, inactive and temporary call signs were filtered. Only 100 active officers with plenty of continuous and frequent GPS records were chosen as the objects of the pilot study.

4.2. *ST-ROIs in foot patrol activity*

Just as places can have different meanings to people at different times, a similar situation also applies to police patrol activities. Many ROIs of officers have their own life spans and the patrols in a day are divided into three shifts (i.e. early, late and night shifts), each lasting for about 8 hours to give each officer proper working hours. Therefore, one place can only be meaningful to the officers during their working hours and the officers can go to the same area to perform different tasks at different moments in time.

To detect the high density aggregation of stays in space and time, we applied ST-DBSCAN to the stay points of all officers to detect their common ST-ROIs. In ST-DBSCAN, SMRD is set to be 60, which is the mean of estimated GPS error of APLS. TMRD is set to be 5 minutes, which is half of the minimum individual sampling interval. Then, MinPts is set to be 65 based on Equation (1) after SMRD and TMRD are set. With ST-DBSCAN, 28 clusters were detected as ST-ROIs (Figure 6). It can be seen that some ST-ROIs are outside the boundary of Camden, and we will discuss the reason for this in the validation stage. The movement of each officer is then captured as the movements and stays from one ST-ROI to another. Time allocation profiles generated from this representation is then compared pair-wise to group the officers.

4.3. *Segregating officer groups*

Using HC, the entire officer community in Camden was segregated into a taxonomy tree structure visualised in Figure 7, with the identification numbers representing each unique officer. The tree can be cut at certain places according to the condition the researcher define to separate the whole dataset into several clusters. In this research, we use Dunn index as this condition.

To test the performance of our proposed similarity (Equation 4) based upon time allocation to ST-ROIs, we compared the HC results with those generated by using the similarity metric defined only by spatial ROIs (Equation 2). Figure 8 shows the performance comparison using the Dunn index.

The similarity based on only spatial ROIs demonstrates better segregations when the cluster number is less than 4, but it falls below the performance of the proposed metric

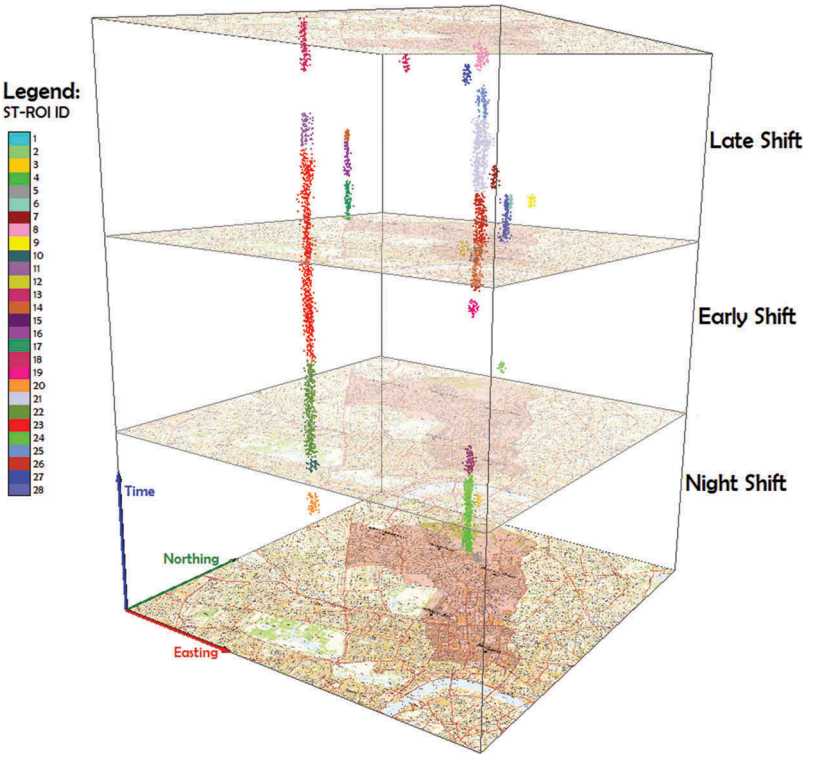

Figure 6. One typical working day of officers is separated into three shifts. 28 ST-ROIs in the case study period of foot patrol officers are detected by ST-DBSCAN and are labelled with different colours.

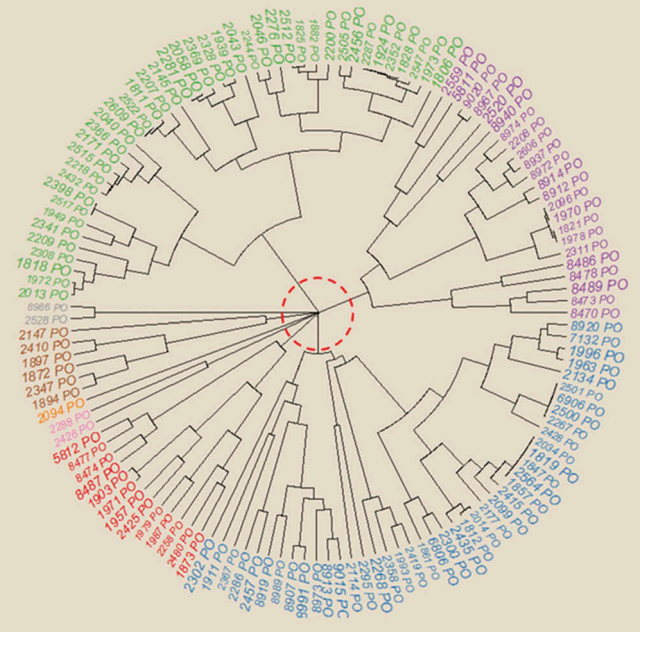

Figure 7. The taxonomy tree showing the clustering results of officers with different patrol patterns. All officers' identification numbers have been encrypted for security reasons

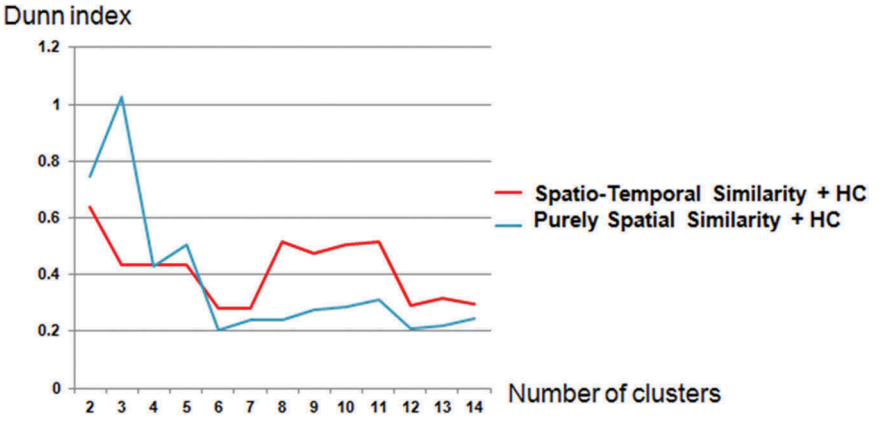

Figure 8. Evaluation of hierarchical clustering results based on two different similarity metrics.

based on the time allocation to ST-POIs with higher cluster numbers. This is partly because the number of detected spatial ROIs is much less than ST-ROIs and the distribution of each user's number of visits to each ROI is therefore much simpler and adapted to segregation of lower cluster numbers. However, a small-cluster-number clustering is not appropriate for semantic explanations of behaviours since a binary or ternary segregation will separate people into groups that are too simple to make sense. For instance, if the officers are only separated into one group that is active inside the Camden border and another group outside, much potential valuable information will be subsumed. According to the cluster number determination method proposed by Salvador and Chan (2004) and the Dunn index evaluation, the number of officer subgroups generated by the HC based on the proposed similarity metric is set to be 8. The red dashed circle in Figure 7 is the place where the tree was cut so that the officers are segregated into eight subgroups.

4.4. *Semantic validation*

It should be noted that effective segregation of the data does not necessarily indicate that the result will make sense in terms of having a reasonable semantic interpretation. To discover the semantic meaning of the generated cluster of space-time profiles, additional information and further study is required. By pinpointing the stay points of each cluster of officers and associating them with building and land use information, the semantic meanings of these differences are revealed. For security reason, we cannot present all the eight clustered officer subgroups, although four of them were randomly chosen as examples to demonstrate the results. Figure 9 shows the mean time percentage allocation to 28 ST-ROIs of the 4 chosen officer subgroups, subgroups I, II, III and IV. Each column in the histograms represents the percentage of time once officer subgroup has spent on one corresponding ST-ROI.

For a more direct and concrete understanding of the discovered differences between the time allocation patterns of the subgroups, the stay points of the four example subgroups are visualised in space-time cubes in Figure 10 to show how different

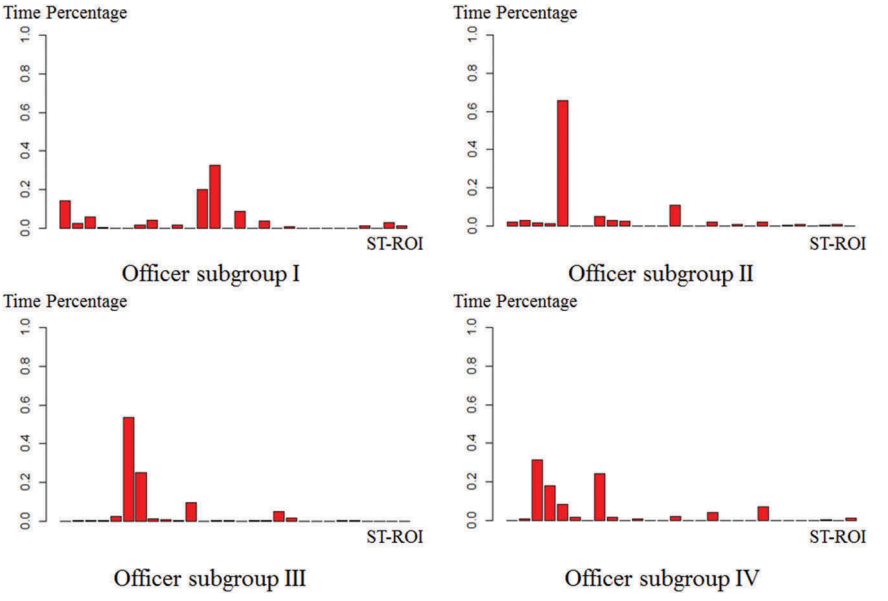

Figure 9. Histograms of mean space-time profiles of different officer subgroups.

subgroups behave differently in both space and time. The base maps in Figure 10 depict the boundary and wards of Camden and the circles marked with numbers and dashed rims on the base maps indicate the spatial location of the stay points on the base map. By associating them with public points of interest data provided by the Edina Digimap, we identified Place No. 1 as an underground station on a commercial streets and Place No. 3 as an underground station in residential area. The same location (Place No. 2) that exists in all the three graphs is the centre of Camden, which is a place with bars, restaurants, a large market and a busy London Underground station. It is a highly populated area and many crimes occur here. Some foot patrol officers spend a long time in Camden town centre because they believe that high visibility has a positive impact on public confidence and acts as deterrence to potential criminals. Place No. 4 is Belgrave Square, an embassy area outside Camden's border, the below mentioned Syrian embassy is located in this area.

It can be seen in Figure 10 that subgroup I has a special interest in places No. 1, No. 3 and sometimes No. 2 during the afternoon peak periods. All of the three places have underground stations inside. The interpretation of this behaviour pattern is that some officers are assigned to focus on peak locations at peak times for high visibility and crime reduction and London Underground stations are often their typical targets.

The aggregation of police force, especially officer subgroup III in this study period, February 2012, was confirmed by the news that hundreds of violent protestors trying to get into the Syrian embassy clashed with the police, and that the police arrested several protestor overnight (Daily Mirror News 2012). It was also explained by the metropolitan police that when there are big events in neighbouring boroughs and extra man power is needed, officers may be ordered to go out of their own borough to help. We can see that

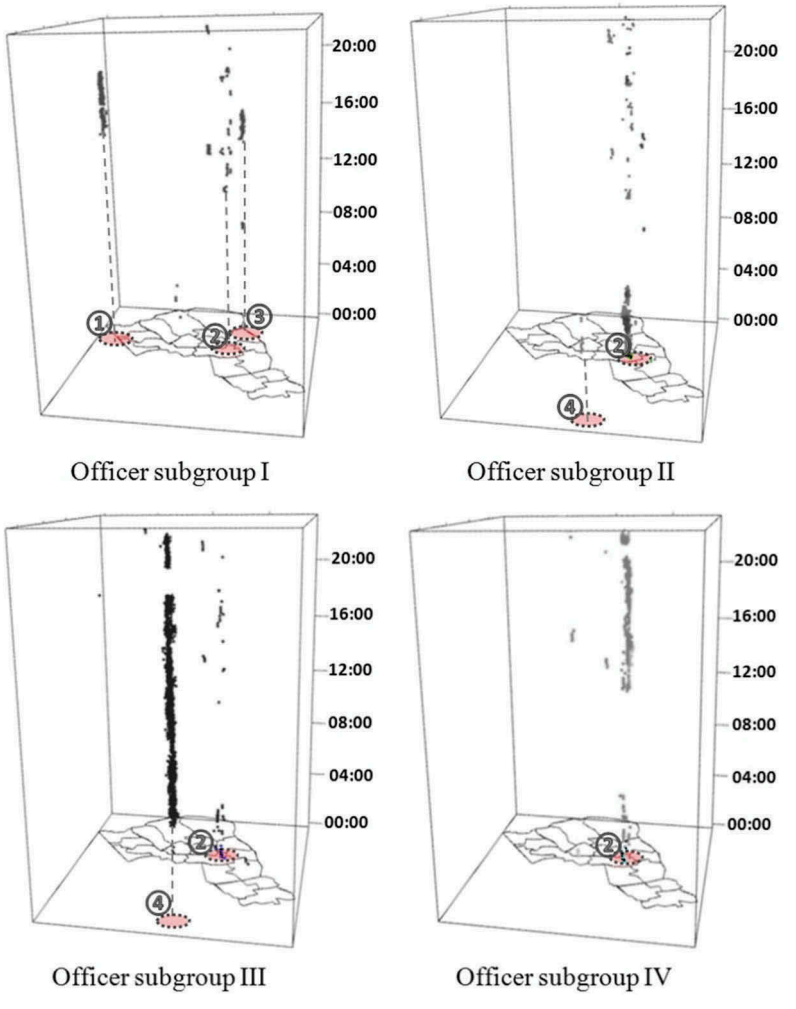

Figure 10. The stay points of four chosen generated officer subgroups. (① Underground station; ② Central Camden; ③ Underground station; ④ Syrian embassy in Belgrave Square).

officer subgroup III spent most of their stay time from the early morning throughout the day in this embassy area while other subgroups spent very little, if any, time there.

It is noteworthy that officer subgroups II and IV were both interested in Place No. 2, the centre of Camden Town. Existing methods that are solely based on location history will not distinguish between them. However, the proposed similarity metrics still managed to distinguish these officers as two groups because their time of visit, length of stay and visit intensity to this common place differed. Officer subgroup II tended to pay frequent visit to central Camden Town at the beginning of the day from 00:00 to 04:00. This phenomenon can be explained by officers keeping an eye on alcohol related and recreational activities in this area at night. In contrast, officer subgroup IV preferred to appear at this area in the afternoon and to stay for a longer time for a different purpose, namely, to maintain a visible presence in an area with large flows of citizens visiting the

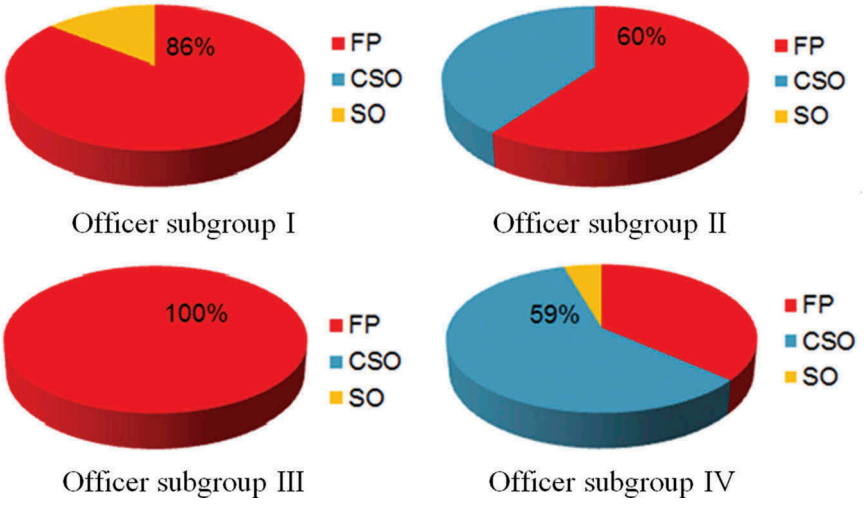

Figure 11. The stay points of four chosen generated officer subgroups.

underground station and shops. Similar analyses can be carried out to explain the patterns of the other subgroups.

Besides, the information of the officer types contained in the APLS dataset can be used for validation as well. Most of the officers on duty are foot patrol officers (FP), community support officers (CSO) and senior officers (SO). The behaviours of different types of officers can be very different because different tasks they are asked to undertake are determined by their types. Figure 11 shows the percentages of these three officer types in the four generated officer subgroups. Officers in subgroup I focus on multiple ST-ROIs and they are mainly consist of foot patrol officers and senior officers while officers that are temporally seconded outside Camden to assist the security work in the embassy area are all made up with foot patrol officers. This may indicate that FPs are interested in multiple places distributed in Camden and only they can be seconded to do the work outside Camden. The most interesting phenomenon is seen in the comparison between subgroup II and IV again. With the two subgroups concentrate their efforts in the same place (Place No. 2 in central Camden) but different time periods, the contribution percentage of foot patrol officers and community support officers within the two subgroups reversed. This is pointed out by the field expert in the Metropolitan Police that the nature of community CSOs' work is to help the FPs at peak places in peak periods and the CSOs do not have much work at night. Similar analyses can be conducted to explain the patterns of the other subgroups.

5. Discussion – conceptual framework of 'who you are'

The method we presented in this case is to group people of a particular occupation (police), which could be used to study the behaviour of other occupations (such as social carers) or socio-demographic attributes. The whole police force in London has about 32,000 officers. Therefore, when we are given a larger user population and a larger study area, there will be much more discovered ST-ROIs for all users. To make sense of

these ST-ROIs, we can use POI (point of interest) feature-type data to extract the semantic meaning of the ST-ROIs and summarise them into generic region types such as stations, commercial streets, politics-related places, residential areas and restaurants. For example, when the police data of all London Boroughs is to be analysed in the future work, the two underground stations identified as ST-ROIs in Figure 10 during the same late shift can be joint and labelled as one 'late shift station' type and the ST-ROIs at night in central Camden can be labelled as 'night shift town centre' type. In this way, the profiles of individuals will be made of the time allocated to each generic ST-ROIs type, not just each specific ST-ROI. This means that the concept can evolve from 'where (ST-ROIs), when and how long you stay is who you are' to 'what place (ST-ROIs types), when and how long you stay is who you are', so that the precise locations of places are replaced by their meanings and people undertaking similar things in difference locations can be grouped together.

6. Conclusions and future work

New datasets of time-series locations enabled the study of behaviours that traditional behaviour studies cannot proceed due to the lack of advanced status logging approaches. In this research, we use police foot patrol data as a case study. The dataset itself focuses on the people of same occupations. However, this is not a major obstacle for our research in which the main goal is to develop a methodological framework to harnesses technological advances to extract group behavioural information from low-level raw GPS trajectories. The framework further extends the traditional ideas of time budget allocation in behavioural studies and existing spatial-location-based user similarity definitions. It is capable of profiling the activity patterns of people according to both space and time aspects by defining a new moving behavioural similarity metric. The clustering analysis based on this similarity metric explains the semantic meaning of various behaviours more reasonably than competing methods. Our contribution also provides a set of computational and visual techniques to human dynamics researchers who may be interested in the variety of individual moving behaviours and helps location-based businesses better understand the characteristics of their customers.

However, there are limitations in the present state of the prototype methods, which will be the directions of future research. First, the generalisation of ST-ROIs is based on density-based clustering and is very time-consuming. A new searching tree and the parallel computation techniques will be used to optimise the retrieval strategy of ST-DBSCAN and speed up the calculation when even larger movement datasets are given.

Second, the current HC method used to group users has limited ability working on large and noisy datasets. New clustering methods such as OPTICS (Ankerst *et al*. 1999) might be more suitable since they are robust in noisy environments and generate results of hierarchical structures. We will use the parallelised OPTICS (Patwary *et al*. 2013) to replace current algorithm and use the UCL's 'Grace' high performance parallel computation platform to deal with large amount of pair-wise comparison.

Last but not least, when the current method is to be applied to large areas (e.g. a much bigger city with huge amount of interesting places for clustering and grouping), we will look into the semantic analysis of places to summarise all the places into a few generic categories so that the proposed method will be able to detect similar behaviours even though they happened in different locations.

Acknowledgements

This work is part of the project – Crime, Policing and Citizenship (CPC): Space-Time Interactions of Dynamic Networks (www.ucl.ac.uk/cpc), supported by the UK Engineering and Physical Sciences Research Council (EP/J004197/1). The data provided by Metropolitan Police Service (London) is highly appreciated.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The first author's PhD research is funded by the China Scholarship Council (CSC). The CSC is a non-profit institution with legal person status affiliated with the Ministry of Education in China.

ORCID

Tao Cheng http://orcid.org/0000-0002-5503-9813

References

Alvares, O., *et al*., 2007. A model for enriching trajectories with semantic geographical information. *In*: *Proceedings of the 15th annual ACM international symposium on advances in geographic information systems*. Article no. 22. New York: ACM. doi:10.1145/1341012.1341041

An, L., *et al*., 2015. Space–time analysis: concepts, quantitative methods, and future directions. *Annals of the Association of American Geographers*, 105 (5), 891–914. doi:10.1080/00045608.2015.1064510

Andrienko, G., *et al*., 2010. Space, time and visual analytics. *International Journal of Geographical Information Science*, 24 (10), 1577–1600. doi:10.1080/13658816.2010.508043

Andrienko, G., *et al*., 2011. From movement tracks through events to places: extracting and characterizing significant places from mobility data. *In*: *IEEE conference on visual analytics science and technology (VAST)*, 23–28 October Providence, RI. Hoboken, NJ: IEEE, 161–170. doi:10.1109/VAST.2011.6102454

Andrienko, N., *et al*., 2015. Scalable and privacy-respectful interactive discovery of place semantics from human mobility traces. *Information Visualization*, 1–37. doi:10.1177/1473871615581216

Ankerst, M., *et al*., 1999. OPTICS: ordering points to identify the clustering structure. *ACM Sigmod Record*, 28 (2), 49–60. doi:10.1145/304181.304187

Baglioni, M., *et al*., 2009. Towards a semantic interpretation of movement behaviour. *In*: *Proceedings of 12th AGILE, lecture notes in geoinformation and cartography*. Berlin: Springer, 271–288. doi:10.1007/978-3-642-00318-9_14

Birant, D. and Kut, A., 2007. ST-DBSCAN: an algorithm for clustering spatial-temporal data. *Data & Knowledge Engineering*, 60, 208–221. doi:10.1016/j.datak.2006.01.013

Cao, X., Cong, G., and Jensen, C.S., 2010. Mining significant semantic locations from GPS data. *Proceedings of the VLDB Endowment*, 3, 1009–1020. doi:10.14778/1920841

Chapin Jr, S., 1974. *Human activity patterns in the city: things people do in time and in space*. Canada: John Wiley & Sons, Inc.

Chen, L., Özsu, M.T., and Oria, V., 2005. Robust and fast similarity search for moving object trajectories. *In*: *Proceedings of the 2005 ACM SIGMOD international conference on management of data*. New York: ACM, 491–502. doi:10.1145/1066157.1066213

Cortez, M., *et al*., 2014. Quality metrics for optimizing parameters tuning in clustering algorithms for extraction of points of interest in human mobility. *In*: *1st symposium on information management and big data*, Cusco, 14 September.

Cullen, G., 1972. Space, time, and the disruption of behaviours in cities. *Paper submitted to the research group on time budges*, Brussels.

Daily Mirror News, 2012. Syria embassy break-in by protesters leads to six arrests [online]. *Daily Mirror UK News*. Available from: http://www.mirror.co.uk/news/uk-news/syria-embassy-break-in-by-protesters-leads-674577 [Accessed 27 June 2015].

Delafontaine, M., *et al*., 2012. Analysing spatiotemporal sequences in Bluetooth tracking data. *Applied Geography*, 34, 659–668. doi:10.1016/j.apgeog.2012.04.003

Dodge, S., Weibel, R., and Forootan, E., 2009. Revealing the physics of movement: comparing the similarity of movement characteristics of different types of moving objects. *Computers, Environment and Urban Systems*, 33 (6), 419–434. doi:10.1016/j.compenvurbsys.2009.07.008

Dunn, J.C., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3 (3), 32–57. doi:10.1080/01969727308546046

Jankowski, P., *et al*., 2010. Discovering landmark preferences and movement patterns from photo postings. *Transactions in GIS*, 14 (6), 833–852. doi:10.1111/tgis.2010.14.issue-6

Kölbl, R. and Helbing, D., 2003. Energy laws in human travel behaviour. *New Journal of Physics*, 5 (48), 48–48. doi:10.1088/1367-2630/5/1/348

Kuijpers, B. and Vaisman, A., 2007. A data model for moving objects supporting aggregation. *In*: *23rd international conference on data engineering*, 17–20 April Istanbul. IEEE, 546–554. doi:10.1109/ICDEW.2007.4401040

Kullback, S. and Leibler, R.A., 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22 (1), 79–86. doi:10.1214/aoms/1177729694

Kwan, M.-P., 2004. GIS methods in time-geographic research: geocomputation and geovisualization of human activity patterns. *Geografiska Annaler, Series B: Human Geography*, 86 (4), 267–280. doi:10.1111/geob.2004.86.issue-4

Kwan, M.-P., Xiao, N., and Ding, G., 2014. Assessing activity pattern similarity with multidimensional sequence alignment based on a multi-objective optimization evolutionary algorithm. *Geographical Analysis*, 46, 297–320. doi:10.1111/gean.2014.46.issue-3

Lee, I., Cai, G., and Lee, K., 2013. Mining points-of-interest association rules from geo-tagged photos. *In*: *Proceedings of the annual Hawaii international conference on system science*, 7–10 January Wailea. IEEE, 1580–1588. doi:10.1109/HICSS.2013.401

Li, Q., *et al*., 2008. Mining user similarity based on location history. *In*: *Proceedings of the 16th ACM SIGSPATIAL international conference on advances in geographic information systems-GIS '08*. Article No. 34. New York: ACM. doi:10.1145/1463434.1463477

Lin, J., 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37 (1), 145–151. doi:10.1109/18.61115

Long, J.A. and Nelson, T.A., 2013. A review of quantitative methods for movement data. *International Journal of Geographical Information Science*, 27 (2), 292–318. doi:10.1080/13658816.2012.682578

Mckenzie, G., 2014. *Activities in a new city: itinerary recommendation based on user similarity*. Banff, AB: Spatial knowledge and information (SKI).

Nara, A., *et al*., 2011. Surgical workflow monitoring based on trajectory data mining. *In*: *New frontiers in artificial intelligence*. Berlin: Springer, 283–291. doi:10.1007/978-3-642-25655-4_27

Palma, A., *et al*., 2009. A clustering-based approach for discovering interesting places in a single trajectory. *In*: *2nd international conference on intelligent computing technology and automation*. New York: ACM, vol. 3, 429–432. doi:10.1145/1363686.1363886

Patwary, A., *et al*., 2013. Scalable parallel optics data clustering using graph algorithmic techniques. *In*: *International conference for high performance computing, networking, storage and analysis*, 17–22 November Denver, CO. IEEE, 1–12. doi:10.1145/2503210.2503255

Patwary, M., *et al*., 2012. A new scalable parallel DBSCAN algorithm using the disjoint-set data structure. *In*: *International conference for high performance computing, networking, storage and analysis*, 10–16 November Salt Lake City, UT. IEEE, 1–11. doi:10.1109/SC.2012.9

Rajasekaran, S., 2005. Efficient parallel hierarchical clustering algorithms. *IEEE Transactions on Parallel & Distributed Systems*, 16 (6), 497–502. doi:10.1109/TPDS.2005.72

Salvador, S. and Chan, P., 2004. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. *In*: *16th IEEE international conference on tools with artificial intelligence*, 15–17 November. IEEE, 576–584. doi:10.1109/ICTAI.2004.50

Shoval, N. and Isaacson, M., 2007. Sequence alignment as a method for human activity analysis in space and time. *Annals of the Association of American Geographers*, 97 (2), 282–297. doi:10.1111/j.1467-8306.2007.00536.x

Spinsanti, L., Celli, F., and Renso, C., 2010. Where you stop is who you are: understanding people's activities by places visited. *CEUR Workshop Proceedings*, 678, 38–52.

Szalai, A., 1966. Trends in comparative time-budget research. *The American Behavioral Scientists*, 9, 9.

Timmermans, H., Arentze, T., and Joh, C.-H., 2002. Analysing space-time behaviour: new approaches to old problems. *Progress in Human Geography*, 26, 175–190. doi:10.1191/0309132502ph363ra

Tsou, M.H., 2015. Research challenges and opportunities in mapping social media and Big Data. *Cartography and Geographic Information Science*, 42 (1), 70–74. doi:10.1080/15230406.2015.1059251

Vazquez-Prokopec, G.M., *et al*., 2013. Using GPS technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment. *Plos One*, 8 (4), e58802. doi:10.1371/journal.pone.0058802

Wald, I. and Havran, V., 2006. On building fast kd-trees for ray tracing, and on doing that in O (N log N). *In*: *Symposium on interactive ray tracing*, 18–20 September Salt Lake City, UT. IEEE, 61–69. doi:10.1109/RT.2006.280216

Ward Jr, J., et al., 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58 (301), 236–244. doi:10.1080/01621459.1963.10500845

Wilson, C., 2001. Activity patterns of Canadian women: application of ClustalG sequence alignment software. *Transportation Research Record: Journal of the Transportation Research Board*, 1777 (1), 55–67. doi:10.3141/1777-06

Wilson, C., 2007. Experiments with activity pattern classification: alignment versus non-alignment methods. *In*: *International Association for Time Use Research annual meeting*, Washington, DC,1–14 October.

Xiao, X., *et al*., 2010. Finding similar users using category-based location history. *In*: *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, vol. 49. New York: ACM, 442–445. doi:10.1145/1869790.1869857

Ye, M., *et al*., 2011. What you are is when you are: the temporal dimension of feature types in location-based social networks. *In*: *Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information system*. New York: ACM, 102–111. doi:10.1145/2093973.2093989

Zheng, Y., *et al*., 2009. GeoLife2.0: a location-based social networking service. *In*: *10th international conference on mobile data management: systems, services and middleware, IEEE Computer Society*, 18–20 May Taipei. IEEE, 357–358. doi:10.1109/MDM.2009.50

Zheng, Y., 2011. *Computing with spatial trajectories*. New York: Springer Science + Business Media.

Zhou, H., et al., 2012. Research on adaptive parameters determination in DBSCAN algorithm. *Journal of Information & Computational Science*, 9 (7), 1967–1973.